interspeech 2008 incorporating SST 08

Prosodic and Spectral Features within Segment-based Acoustic Modeling

Björn Schuller, Xiaohua Zhang, and Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
D-80333 Mnchen, Germany
schuller@tum.de

Abstract

Apart from the usually employed MFCC, PLP, and energy feature information, also duration, low order formants, pitch, and center-of-gravity-based features are known to carry valuable information for phoneme recognition. This work investigates their individual performance within segment-based acoustic modeling. Also, experiments optimizing a feature space spanned by this set, exclusively, are reported, using CFSS feature space optimization and speaker adaptation. All tests are carried out with SVM on the open IFA-corpus of 47 Dutch hand-labeled phonemes with a total of 178k instances. Extensive speaker dependent vs. independent test-runs are discussed as well as four different speaking styles reaching from informal to formal: informal and retold story telling, and read aloud with fixed and variable content. Results show the potential of these rather uncommon features, as e.g. based on F3 or pitch.

Index Terms: phoneme-recognition, prosodic features, acoustic modeling, feature space optimization, ASR

1. Introduction

In order to advance the performance of today's speech recognition engines, many efforts are undertaken from an architectural point of view. As a certain saturation point seems to have been reached, recently more efforts are also spent on re-investigation of features. Apart from the predominant MFCC and PLP, a variety of prosodic, and other spectral characteristics is not used on a broad basis, though well known to carry information about phonemes. Further, these features are often extracted for other speech analysis purposes as e.g. emotion recognition, or detection of non-verbals [1]: likewise they could be integrated at "no cost" in a system that recognizes e.g. emotion and speech. At the same time it seems interesting how much these such features also model phonetic content.

This paper therefore reports on observed accuracies for diverse feature types on the large Dutch IFA-Corpus of hand-labeled phonemes. We discuss feature type relevance and the impact of speaking style on the overall accuracy.

The paper is structured as follows: sect. 2 introduces the used IFA-corpus used. Next, sect. 3 discusses the feature set considered for experiments. In sect. 4 and sect. 5 classification and feature selection will be described. Finally, extensive experimental results are presented in sect. 6, and discussed in sect. 7.

2. IFA-Corpus

In this work we decided for a hand-segmented and hand-labeled corpus on the phoneme level, to minimize statistical noise deriving from mis-alignment. The public IFA corpus [2] of manually transcribed Dutch speech seems a good choice. It consists of 18 speakers (9 male and 9 female) of which 8 speakers (4 male and 4 female, 15 to 66 years of age) were chosen for phonemic segmentation of 51782 segmented words, 187544 segmented phonemes based on complete recordings within 47 phoneme classes. It provides two-channel recordings: a head-mounted dynamic microphone and a fixed HF condenser microphone. Again, we chose the head-mounted recording to minimize other influences such as room acoustics. Eight speaking styles reaching from informal to very formal are contained. These will be ignored in first tests. However, later on we will provide separate results for these speaking styles. In this set the larger subset is marked as "valid" with respect to labeling. Only these instances will be used in the ongoing. Results for data-driven experiments can be found e.g. in [3].

3. Prosodic and Spectral Features

A number of acoustic features is provided for this corpus and tests to be run. For our experiments we consider the set of 6+4 Low-Level-Descriptors (LLD) and functionals, as depicted in table 1 [2]. LLD thereby cover prosodic feature types, namely pitch and intensity, as well as spectral, that is low order formant positions, and center-of-gravity (COG). As segmentation is provided, functionals such as values at five different relative time intervals can be used. Further, minimum, maximum, mean and standard deviance are contained. Note that no delta coeffi-

Accepted after peer review of full paper

LLD	Functional
C Center-of-Gravity	*Min* Min. value inside seg.
I Intensity	*Max* Max. value inside seg.
$F1$ Formant Position	*Time0* Value at start of seg.
$F2$ Formant Position	*Time1* Value 1/4 of seg.
$F3$ Formant Position	*Time2* Value at center of seg.
P Pitch	*Time3* Value 3/4 of seg.
Smooth F1 Pos.	*Time4* Value at end of seg.
Smooth F2 Pos.	*Tmax* Rel. time of Max. (0-4)
Smooth F3 Pos.	*Tmin* Rel. time of Min. (0-4)
Smooth Pitch	*Mean* Mean Value (F0 only)
	StdDev Stand. Dev. (F0 only)

Table 1: *Low-Level-Descriptor (LLD) and functional types used for feature space construction.*

cients are added and pitch and formants exist in two versions: a smoothed version by global view and DP vs. a raw individually frame by frame extracted version. In total, 92 static features per segment result from application of functionals to the LLD.

4. Classification

Herein we do not employ Gaussian Mixtures for classification of phonemes, as would usually be the case in an HMM framework. We rather use Support Vector Machines (SVM) with polynomial Kernel, pairwise multiple class discrimination, and Sequential Minimal Optimization (SMO) as introduced in [4]. This choice derives from the fact that we use a larger feature space than in usual phoneme recognition, which mostly bases on a 39-dimensional feature vector based upon first 12 MFCC and frame energy plus speed and acceleration coefficients. As opposed to this, our feature vector consists of the 92 features, as introduced in 3, which is intended to be enlarged in future applications, where cepstral coefficients will be added. Further, we usually employ hybrid architectures combining discriminative abilities of Neural Nets [5] or SVM [6] with the warping capabilities of HMM. This also motivates provision of results based on SVM.

5. Feature Selection

Next, we consider removal of irrelevant and redundant information, as it often improves the performance of machine learning algorithms. Such feature selection retains only a subset of the original features, and reveals relevance of features - ideally, the set which gives the best possible classification accuracy. Exhaustive search is computationally prohibitive in our case, except for a small number of dimensions, as it would involve generating and testing $2^n - 1$ possible combinations with n being the dimension of the feature space. Likewise we are forced to search for suboptimal solutions. In general, there are two approaches to feature selection: filter and wrapper methods. Open-loop filter methods do not use classifier feedback to determine best features. In this group, well known Correlation-based Feature Subset Selection (CFSS) or Information-Gain-Ratio Attribute Evaluation (IGR) and Principal Components Analysis (PCA) are found. As opposed to this, in closed-loop wrapper methods, the classifier's error serves as target function combined with a search function. Best known, Sequential Forward Floating Selection (SFFS) belongs to this group besides e.g. genetic or random search. We decided for a de-correlation by CFSS that optimizes a feature set rather than finding individually relevant features as IGR, and likewise achieves very high compression rates at high accuracy levels. At the same time it is a fast selection. At its heart is a heuristic for evaluating the worth or merit of a subset of features. This heuristic takes into account the usefulness of individual features for predicting the class label along with the level of inter correlation among them [7]. The hypothesis behind this heuristic is: good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. In test theory the same principle is used to design a composite test for predicting an external variable of interest. In this situation the features are individual tests which measure traits related to the variable of interest (class) by Pearson's correlation where all variables have been standardized.

CFSS first calculates a matrix of feature-class and feature-feature correlations from the training data, and then searches the feature subset space using e.g. a best first search, as herein. This search starts in a hill-climbing manner with an empty set of features and generates all possible single feature expansions. The subset with the highest evaluation is chosen and expanded in the same manner by adding single features. If expanding a subset results in no improvement, the search drops back to the next best unexpanded subset and continues from there. Given enough time, a best first search will explore the entire feature subset space, so in common to limit the number of subsets expanded that result in no improvement. The best subset found is returned when the search terminates. We use a stopping criterion of five consecutive fully expanded no-improving subsets.

6. Experimental Results

To provide results for speaker-dependent and independent evaluation, we employ j-fold stratified cross validation (SCV) and leave-one-speaker-out (LOSO). This allows for training disjunctive results on the whole corpus. We report mean accuracies of correctly assigned phonemes vs. all phonemes.

In table 2 results for speaker dependent analysis are shown for each of the 4 female speakers indexed as F20N, F28G, F40L, and F60E and the four male speakers indexed as M15R, M40K, M56H, and M66O in [8]. The

Speaker ID	**F20N**	**F28G**	**F40L**	**F60E**
Features [#]	30	36	30	32
Instances [#]	24999	35243	24452	31440
Accuracy [%]	57.02	65.34	59.84	59.36
Mean Acc. [%]	60.77			
Speaker ID	**M15R**	**M40K**	**M56H**	**M66O**
Features [#]	25	35	31	35
Instances [#]	15706	15798	21747	9238
Accuracy [%]	49.45	52.80	52.98	61.60
Mean Acc. [%]	53.22			
Total Acc. [%]	**58.16**			

Table 2: *Accuracy per speaker, speaker-dependent analysis, CFSS, IFA-Corpus, 178623 phonemes, SVM, 3-fold SCV. If MFCC plus Δ and $\Delta\Delta$ are used with all functionals as shown in table 1 under same conditions, 75.66% total accuracy are obtained (76.87% mean accuracy female; 73.42% male speakers).*

table shows the number of valid instances contained for these speakers in the IFA corpus. At this point all material is used, independent of the speaking style. Note that notably fewer instances are available for the male speakers. Also, accuracy per phoneme is significantly lower for them.

Next, in table 3, we show the same results for speaker independent analysis. A LOSO evaluation per gender was used, here, and reveals a clear downgrade in accuracy of almost 10% absolute. No positive effect could be obtained by speaker normalization to zero mean and ± 1 standard deviance with use of the whole speaker context for each speaker. Again, female speakers' phonemes are recognized with higher accuracy, probably deriving from the higher amounts of available data. Feature space optimization was carried out only once per gender rather than individually per speaker.

Speaker ID	**F20N**	**F28G**	**F40L**	**F60E**
Features [#]	29			
Instances [#]	24999	35243	24452	31440
Accuracy [%]	50.51	53.25	54.20	51.17
Mean Acc. [%]	52.30			
Speaker ID	**M15R**	**M40K**	**M56H**	**M66O**
Features [#]	30			
Instances [#]	15706	15798	21747	9238
Accuracy [%]	39.79	40.76	46.29	53.45
Mean Acc. [%]	44.32			
Total Acc. [%]	**49.51**			

Table 3: *Accuracy per speaker, speaker-independent analysis, CFSS, IFA-Corpus, 178623 phonemes, SVM, LOSO.*

The distribution of features found within speaker dependent and independent feature selection is depicted in table 4. No significant difference can be found for female or male speakers and speaker independent vs. speaker dependent recognition. However, raw features are clearly preferred by the selection over their smoothed versions. First come COG and intensity. Then especially F1 and F2 show a high contribution: both by raw and smoothed features. Interestingly, also F3 has a remarkable weight in terms of total number of features. Pitch is among least important as well - as to be expected, yet it seems noteworthy that pitch features *are* selected.

	raw		raw/smooth			
Dim. [#]	**C**	**I**	**F1**	**F2**	**F3**	**P**
	f					
sd	7	5	4/4	5/3	2/2	2/0
si	7	5	5/2	5/2	1/0	2/0
	m					
sd	8	5	4/4	5/1	1/2	1/1
si	8	5	3/4	5/0	2/1	1/1

Table 4: *Distribution of features: speaker dependent (sd) and independent (si) analysis, CFSS, IFA-Corpus, 178623 phonemes, SVM, 3-fold SCV/LOSO.*

Next, we consider effect of diverse speaking styles: Table 5 reveals the influence of speaking style on the overall accuracy. Of the original eight categories as named in [8], we use only four by clustering the read aloud styles into such with variable content and such with fixed content as opposed to the informal styles story telling with sight contact to an interviewer, and story retelling without sight contact. A clear impact on accuracy can be observed, as would also be expected: from informal to formal accuracy raises within 10-fold SCV by 15% absolute, thus clearly stressing the difficulty of informal speech handling.

Style	**inform.**	**retold**	**read vc**	**read fc**
Gender	**f**			
Features [#]	28	31	29	33
Instances [#]	14627	11847	46679	42981
Accuracy [%]	47.10	53.52	58.65	61.78
Gender	**m**			
Features [#]	16	26	29	32
Instances [#]	2917	8694	14917	32961
Accuracy [%]	23.89	43.81	49.92	55.32
Total Acc. [%]	**43.24**	**49.41**	**56.54**	**58.98**

Table 5: *Accuracy per speaking style: informal, retold, read aloud with variable (vc) and fixed (fc) content, CFSS, IFA-Corpus, 175623 phonemes, SVM, 10-fold SCV.*

In table 6 we also show feature distribution after CFSS for the different speaking styles. Again, a very constant picture is observed: while speaking style highly in-

fluenced recognition performance, it apparently has only little effect on feature distribution. Gender has more influence at this point, which however may well derive from fewer instances for male speakers, as denoted. The only tendency is a decreasing relevance of intensity and the rather unusual F3 Position with increasing informality.

	raw		raw/smooth			
Dim. [#]	**C**	**I**	**F1**	**F2**	**F3**	**P**
			f			
Informal	7	4	4/2	3/3	1/0	1/1
Retold	7	5	5/2	4/3	1/0	2/0
Read vc	7	5	4/3	4/1	1/0	2/0
Read fc	7	5	5/3	5/2	1/1	2/0
Mean	**7.0**	**4.8**	**7.0**	**6.3**	**1.3**	**2.0**
			m			
Informal	5	2	2/1	3/0	0/0	1/0
Retold	7	3	3/3	5/0	1/0	2/0
Read vc	8	5	3/3	5/0	1/0	1/1
Read fc	8	5	4/3	5/0	1/2	1/1
Mean	**7.0**	**3.8**	**5.5**	**4.5**	**1.3**	**1.8**

Table 6: *Distribution of features per speaking style: informal, retold, read aloud with variable (vc) and fixed (fc) content, CFSS, IFA-Corpus, 175623 phonemes, SVM, 10-fold SCV.*

7. Conclusion and Outlook

In this paper we showed results for recognition of pre-segmented phonemes on the IFA corpus with non-cepstral features as alternatives for acoustic modeling. Throughout experiments no "real surprise" was observed with respect to feature importance: center of gravity comes first, followed by intensity, and formant positions 1-3, while formant 3 is almost least. Its non-the-less measurable importance may thereby arguably derive from the extra contrasting category of front rounded vowels in Dutch, as F3 is needed to distinguish this natural class from the [-round, -back] vowels.

Last comes pitch, yet it still contributes to the phoneme recognition task. Thereby the question whether this is due to its prosodic, or its non-prosodic function qua underlying the contrast between [voiced] and [voiceless] segments. Same applies mutatis mutandis for intensity presumably being a good discriminator between [+/- consonantal] segments, and within the [+cons] category between [+/- sonorant] segments.

In general raw feature variants are preferred over smoothed ones.

Roughly 60% accuracy can be reported employing only these non-cepstral features, already, as opposed to the usual rates around 80% plus by MFCC [9]. Considering speaking style, expectancy is also fulfilled: best accuracy is observed for the formal reading aloud of fixed content, followed by variable content, and more informal styles retelling or freely telling stories. However, the results show the generally high capability of prosodic and spectral features for the discrimination of phonemes already "stand-alone".

In future works we aim at inclusion of the features found most relevant in an SVM/HMM ASR framework.

8. References

[1] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals," in *Proc. INTERSPEECH 2007*, Antwerp, Belgium, 2007, pp. 2253–2256.

[2] R. van Son, D. Binnenpoorte, H. van den Heuvel, and L. Pols, "The ifa corpus: a phonemically segmented dutch 'open source' speech database," in *Proc. of Eurospeech 2001*, 2001.

[3] L. Bosch, "Speech variation and the use of distance metrics on the articulatory feature space.," in *Proc. ITRW Workshop on Speech Recognition and Intrinsic Variation*, 2006.

[4] J.C. Platt, "Sequential minimal optimization: A fast algrithm for training suport vector machine.," in *Microsoft Reasearch Tech. Teport MSR-TR-98-14, Microsoft, Redmond*, 1998.

[5] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation.," in *Proc. ISCA Speech Prosody 2006*. ISCA, 2006.

[6] J. Stadermann and G. Rigoll, "A hybrid svm/hmm acoustic modeling approach to automatic speech recognition.," in *Proc. Interspeech 2004 ICSLP 2004*. ISCA, 2004, vol. I, pp. 661–664.

[7] M. A. Hall, "Correlation-based feature selection for machine learning.," in *PhD diss. Hamilton, NZ: Waikato University, Department of Computer Science*, 1998.

[8] R. van Son and L. Pols, "Structure and access of the open source ifa corpus," in *Proc. of the IRCS workshop on Linguistic Databases*, 2001, pp. 245–253.

[9] L. Bosch, H. Baayen, and M. Ernestus, "On speech variation and word type differentiation by articulatory feature," in *Proc. Interspeech 2006*. ISCA, 2006.